TALN 2010, Montréal, 19–23 juillet 2010

Évaluation automatique de résumés avec et sans référence

Juan-Manuel Torres-Moreno[1, 2] Horacio Saggion[3] Iria da Cunha[1, 4]
Patricia Velázquez-Morales[5] Eric SanJuan[1]
(1) LIA, Université d'Avignon et des Pays de Vaucluse, Avignon, France
(2) Ecole Polytechnique de Montréal, (Québec) Canada
(3) DTIC, Universtitat Pompeu Fabra, Barcelona, Espagne
(4) IULA, Universitat Pompeu Fabra, Barcelona, Espagne
(5) VM Labs, Avignon, France
juan-manuel.torres@univ-avignon.fr, {horacio.saggion, iria.dacunha}@upf.edu

Résumé. Nous étudions différentes méthodes d'évaluation de résumé de documents basées sur le contenu. Nous nous intéressons en particulier à la corrélation entre les mesures d'évaluation avec et sans référence humaine. Nous avons développé FRESA, un nouveau système d'évaluation fondé sur le contenu qui calcule les divergences entre les distributions de probabilité. Nous appliquons notre système de comparaison aux diverses mesures d'évaluation bien connues en résumé de texte telles que la Couverture, *Responsiveness*, *Pyramids* et *Rouge* en étudiant leurs associations dans les tâches du résumé multi-document générique (francais/anglais), focalisé (anglais) et résumé mono-document générique (français/espagnol).

Abstract. We study document-summary content-based evaluation methods in text summarization and we investigate the correlation among evaluation measures with and without human models. We apply our comparison framework to various well-established content-based evaluation measures in text summarization such as Coverage, *Responsiveness*, *Pyramids* and *Rouge* studying their associations in various text summarization tasks including generic (English/French) and focus-based (English) multi-document summarization and generic multi and single-document summarization (French/Spanish). The research is carried out using the new content-based evaluation framework FRESA to compute the divergences among probability distributions.

Mots-clés : Mesures d'évaluation, Résumé automatique de textes.

Keywords: Evaluation measures, Text Automatic Summarization.

1 Introduction

L'évaluation des résumés produits de manière automatique a toujours été une question complexe et controversée du Traitement Automatique de la Langue (TAL). Au cours des dernières années, des évaluations à grande échelle, indépendantes des concepteurs des systèmes, ont vu le jour et plusieurs mesures d'évaluation ont été proposées. En ce qui concerne l'évaluation des systèmes de résumé automatique, deux campagnes d'évaluation ont déjà été menées par l'agence américaine DARPA (*Defense Advanced Research Projects Agency*). La première, intitulée SUMMAC, s'est déroulée de 1996 à 1998 sous l'égide du programme TIPSTER (Mani *et al.*, 2002), et la deuxième, intitulée DUC (*Document Understanding Conferences*) (Over *et al.*, 2007) a suivi de 2000 á 2007. Depuis 2008 c'est la *Text Analysis Conference*

(TAC) (TAC, 2008) qui a pris la suite et est le forum pour l'évaluation des différentes technologies d'accès à l'information textuelle, y compris le résumé de texte. Les évaluations du résumé de texte sont de deux types : extrinsèque et intrinsèque (Spärck Jones & Galliers, 1996). Dans une évaluation extrinsèque, les résumés sont évalués dans le contexte d'une tâche spécifique réalisée par un humain ou une machine. Dans l'évaluation intrinsèque, les résumés sont évalués par rapport à une référence ou modèle idéal. SUMMAC a suivi un paradigme d'évaluation extrinsèque et DUC/TAC ont suivi celui de l'évaluation intrinsèque. Afin d'évaluer intrinsèquement les résumés, un résumé candidat (*peer*) est comparé à un ou plusieurs résumés de référence (*models*). DUC a utilisé l'interface SEE qui permet aux juges humains de comparer un *peer* aux *models*. Les juges attribuent ainsi une note de couverture au résumé candidat et la note finale (score) est la moyenne des notes obtenues. Le score final peut alors être utilisé pour établir un classement (*ranking*) des systèmes de résumé automatique (résumeurs). Dans le cas du résumé orienté par une requête (par exemple, lorsque le résumé doit répondre à une ou à un ensemble de questions) un score d'adéquation de la réponse *Responsiveness* est assigné au résumé. Ce score évalue la manière dont le résumé répond aux questions. Puisque la comparaison manuelle des résumés candidats avec ceux de référence est un processus ardu et coûteux, des recherches ont été menées ces dernières années sur les procédures d'évaluation automatique fondées sur le contenu. Les premières études utilisaient des mesures de similarité telles que le cosinus (avec ou sans pondération) pour comparer les résumés candidats et ceux de référence (Donaway *et al.*, 2000). Diverses mesures de Couverture du vocabulaire telles que les n-grammes ou la plus longue sous-séquence commune entre le candidat et le modèle ont été proposées (Radev *et al.*, 2003). La mesure d'évaluation des systèmes de traduction automatique BLEU (Papineni *et al.*, 2002) a été également testée en résumé de texte par Pastra & Saggion (2003). Les conférences DUC ont adopté ROUGE (Lin, 2004) pour l'évaluation du résumé fondée sur le contenu. Il a été montré que les classements des systèmes générés par certaines mesures ROUGE (par exemple, ROUGE-2 qui utilise 2-grammes) ont une bonne corrélation avec les classements produits par couverture. Ces dernières années, la méthode d'évaluation PYRAMIDS a été introduite par Nenkova & Passonneau (2004). PYRAMIDS est basée sur la distribution de l'informativité (*content*) d'un ensemble de résumés de référence. Les unités informatives (*Summary Content Units*, SCU) sont d'abord identifiées dans les résumés de référence, puis chacune reçoit un poids égal au nombre de références contenant la même unité. Les SCU des résumés candidats identifiées sont alignées contre celles des références, puis pondérées. Le score PYRAMIDS attribué aux candidats est le rapport entre la somme des poids de ces unités et la somme des poids du meilleur résumé idéale possible avec le même nombre d'unités SCUs des candidats. Les scores PYRAMIDS sont aussi utilisés pour le classement des systèmes. Nenkova & Passonneau (2004) ont montré que les scores PYRAMIDS produisent de classements fiables lorsque plusieurs (4 ou plus) références sont utilisées et que les classements PYRAMIDS sont en corrélation avec ceux produits par ROUGE-2 et ROUGE-SU4 (ROUGE avec *skip* 2-grammes). Cependant cette méthode nécessite la création de références et l'identification, l'alignement et la pondération des SCU dans les références et dans les candidats. Nenkova & Passonneau (2004) ont proposé d'utiliser directement le document complet à des fins de comparaison et ont argumenté que les mesures basées sur le contenu, qui comparent le document entier au résumé, pourraient être des substituts acceptables à celles utilisant les résumés de référence. Une nouvelle méthode d'évaluation de systèmes de résumé sans référence a été récemment proposée (Louis & Nenkova, 2009). Elle est basée sur la comparaison directe de l'information contenue entre les résumés et leurs documents. Louis & Nenkova (2009) ont évalué l'efficacité de la mesure théorique de Jensen-Shannon ($\mathcal{JS}$) (Lin, 1991) pour le classement des systèmes dans les tâches de résumé multi-document focalisé sur une requête. Elles ont montré que les classements produits par PYRAMIDS et ceux produits par la mesure $\mathcal{JS}$ sont corrélés, même si on ne tenait pas compte de la requête dans l'évaluation, c'est à dire que l'on ramène en première approximation la tâche de résumé guidé à celle de résumé générique. On peut cependant étudier l'effet de la mesure $\mathcal{JS}$ dans une réelle tâche de

résumé générique multi-document en se référent à la tâche 2 de DUC'04, ce que nous faisons ici. Nous nous intéressons aussi à d'autres types de résumés tels que le résumé biographique (DUC'04 tâche 5), le résumé d'opinions (TAC'08 OS) et le résumé dans de langues autres que l'anglais. Dans cet article, nous présentons une série d'expériences visant une meilleure compréhension de la pertinence de la divergence $\mathcal{JS}$ pour le classement des systèmes de résumé. Nous avons effectué des expériences avec la mesure $\mathcal{JS}$ et nous avons vérifié que, pour certaines tâches (telles que celles étudiées par Louis & Nenkova (2009) il existe une forte corrélation entre PYRAMIDS, *Responsiveness* et la divergence $\mathcal{JS}$, mais comme nous allons le montrer plus loin, il existe aussi des jeux de données de référence pour lesquelles la corrélation n'est pas si forte. Nous présentons aussi des expériences sur des jeux de données en espagnol et en français qui montrent aussi une corrélation positive entre les mesures $\mathcal{JS}$ et ROUGE. Cet article est organisé de la façon suivante : dans la section 2 nous présentons les travaux existants dans le domaine de l'évaluation du résumé basée sur le contenu, ce qui nous permet de préciser le point de départ de nos recherches. Dans la section 3, nous décrivons la méthodologie suivie, les outils ainsi que les ressources utilisés lors des expériences. Dans la section 4, nous présentons les expériences menées et les résultats obtenus sur les différents corpora et tâches. En section 5 comprend une discussion sur ces résultats, avant de conclure et de présenter quelques perspectives et travaux futurs.

2 Etat de l'art

Donaway *et al.* (2000) est un des premiers travaux qui mentionne l'utilisation de mesures fondées sur le contenu. Il a présenté une méthode d'évaluation des systèmes de résumé automatique au moyen de calculs de rappel et de variantes de la distance du cosinus entre le texte et le résumé produit. Ce dernier calcul est clairement une mesure fondée sur le contenu. Ils ont montré qu'il y avait une corrélation faible entre les classements produits par le calcul du rappel, mais que les mesures basées sur le contenu produisaient des classements fortement corrélés. Ceci a ouvert la voie aux mesures fondées sur le contenu, en comparant le contenu du résumé automatique à ceux des résumés de référence. Saggion *et al.* (2002) ont présenté un ensemble de mesures d'évaluation basées sur la notion de chevauchement du vocabulaire qui inclut les n-grammes, la similarité cosinus et la plus longue sous-séquence commune. Ils les ont appliquées au résumé automatique multi-document en anglais et en chinois. Toutefois, ils n'ont pas évalué les performances de ces mesures sur différentes tâches de résumé. Radev *et al.* (2003) ont également comparé différentes mesures d'évaluation basées sur le chevauchement du vocabulaire. Bien que ces mesures permettaient de séparer les systèmes aléatoires de ceux non-aléatoires, aucune conclusion claire ne s'est dégagée sur la pertinence des mesures étudiées. Un système d'évaluation de résumé bien répandu est ROUGE, qui offre un ensemble de statistiques pour comparer les résumés candidats avec un ensemble de références produites par des experts. Diverses statistiques existent selon le n-gramme utilisé et le type de traitement appliqué aux textes d'entrée (par exemple lemmatisation, suppression de mots fonctionnels). Lin *et al.* (2006) ont proposé une méthode d'évaluation basée sur l'utilisation de mesures de divergence entre deux distributions de probabilité (la distribution d'unités dans le résumé automatique et celle d'unités dans le résumé de référence). Ils ont étudié deux mesures théoriques d'information : la divergence de Kullback-Leibler ($\mathcal{KL}$) (Kullback & Leibler, 1951) et celle de Jensen-Shannon ($\mathcal{JS}$) (Lin, 1991). La divergence $\mathcal{JS}$ est définie par :

$$D_{\mathcal{JS}}(P||Q) = \frac{1}{2}\sum_{w} P_w \log_2 \frac{2P_w}{P_w + Q_w} + Q_w \log_2 \frac{2Q_w}{P_w + Q_w} \quad (1)$$

Ces mesures peuvent être appliquées à la distribution d'unités dans les résumés des systèmes P et dans ceux de référence Q. La valeur obtenue est utilisée pour attribuer un score au résumé produit. La méthode a été évaluée par Lin *et al.* (2006) sur le corpus DUC'02, pour les tâches de résumé mono et multi-document. Une bonne corrélation a été trouvée entre les mesures de divergence et les deux classements obtenus avec ROUGE et la couverture. Louis & Nenkova (2009) sont allées encore plus loin et, comme Donaway *et al.* (2000), ont proposé de comparer directement la distribution de mots dans les documents complets avec celle des mots dans les résumés automatiques afin d'inférer une mesure d'évaluation basée sur le contenu. Elles ont constaté une forte corrélation entre les classements produits avec références et ceux obtenus sans référence. Ce travail est le point de départ de nos recherches sur la pertinence des mesures qui ne reposent pas sur des références humaines.

3 Protocole d'étude

La méthodologie suivie dans cet article reflète celle adoptée dans les travaux passés (Donaway *et al.*, 2000; Radev *et al.*, 2003; Louis & Nenkova, 2009). Étant donné une tâche de résumé spécifique T, une jeu de p spécifications textuelles :$\{I_i\}_{i=0}^{p-1}$ (par exemple, document(s), question(s), topics) devant guider les résumés produits, s résumés candidats $\{\text{SUM}_{i,k}\}_{k=0}^{s}$ par entrée i, et m résumés de référence $\{\text{REF}_{i,j}\}_{j=0}^{m}$ par entrée i, nous allons comparer les classements produits par différentes mesures d'évaluation basées sur le contenu, de s systèmes de résumé automatique. Certaines mesures sont utilisées pour comparer les résumés automatiques avec n des m références humaines :

$$\text{MESURE}_M(\text{SUM}_{i,k}, \{\text{REF}_{i,j}\}_{j=0}^{n}) \tag{2}$$

tandis que d'autres mesures comparent les résumés candidats avec l'entrée ou une partie de l'entrée :

$$\text{MESURE}_M(\text{SUM}_{i,k}, I'_i) \tag{3}$$

où I'_i est un sous-ensemble des entrées I_i. On obtient la moyenne des valeurs produites par les mesures pour chaque résumé $\text{SUM}_{i,k}$ et pour chaque système $k = 0, \ldots, s$. Ces moyennes induisent un classement. Ensuite, les classements sont comparés avec le taux ρ de corrélation de Spearman (Siegel & Castellan, 1998), utilisé pour mesurer le degré d'association entre deux variables dont les valeurs servent à classer des objets. Nous avons choisi d'utiliser cette corrélation pour comparer directement les résultats à ceux présentés par Louis & Nenkova (2009). Les calculs des corrélations ont été effectués avec le logiciel *Statistics-RankCorrelation-0.12*[1], qui calcule la corrélation du classement entre deux vecteurs. Nous avons par ailleurs vérifié la bonne conformité de ces résultats avec le test de corrélation du τ de Kendall calculé avec le logiciel de statistique R. Les deux tests non paramétriques de Spearman et de Kendall ne se distinguent réellement que sur le traitement des ex-æquo. La bonne correspondance entre les deux tests montre que ces derniers n'introduisent pas de biais dans nos analyses. Par la suite nous ne mentionneront que le ρ de Sperman, plus largement utilisé dans ce domaine.

3.1 Outils

Les expériences ont été réalisées en utilisant un nouveau système d'évaluation de résumés : FRESA – *FRamework for Evaluating Summaries Automatically*– qui inclut des mesures d'évaluation fondées sur

[1]CPAN, `http://search.cpan.org/~gene/Statistics-RankCorrelation-0.12/`

la distribution de probabilités. De façon similaire à ROUGE, FRESA utilise des n-grammes et des *skip* n-grammes dans le calcul des distributions de probabilité. L'environnement FRESA peut être utilisé pour l'évaluation de résumés en anglais, français, espagnol et catalan. Il intègre filtrage et lemmatisation dans le pre-traitement des documents. Il a été développé en Perl et mis à disposition de la communauté[2]. Nous utilisons aussi ROUGE pour calculer diverses statistiques des corpora de test.

3.2 Tâches de résumé et corpora

Nous avons mené nos expériences sur les corpus et les tâches de résumé suivants :

1. Résumé générique multi-document en anglais (génération d'un petit résumé à partir d'un groupe de documents pertinents d'une thématique donnée) de DUC'04[3], tâche 2 : 50 groupes, 10 documents par groupe, 294.636 mots ;
2. Résumé guidé (*Focused-based*) en anglais (par exemple, génération d'un résumé multi-document guidé par la question "qui est X ?", où X est le nom d'une personne) à partir des données DUC'04, tâche 5 : 50 groupes, 10 documents dans chaque plus le nom d'une personne-cible, 284.440 mots ;
3. Tâche mise à jour, qui consiste à créer un résumé d'un groupe de documents et une thématique. Deux sous-tâches sont considérées : a) un premier résumé doit être produit à partir d'un ensemble de documents et d'une thématique ; b) une mise à jour du résumé doit être réalisée à partir d'un groupe différent (mais lié) en supposant que les documents utilisés en a) ont été lus. Le corpus de résumés mis à jour TAC'08 en anglais est utilisé : 48 thèmes, 20 documents chacun, 36.911 mots.
4. Résumé d'opinions où les systèmes résument les opinions sur une entité cible dans un ensemble d'articles de blogs. Le corpus TAC'08 OS en anglais[4] (tiré de la collection de textes du Blogs06) a été utilisé : 25 groupes et cibles (par exemple, la cible entité et questions), 1,167.735 mots.
5. Résumé générique mono-document en espagnol, corpus *Medicina Clínica*[5] composé de 50 articles médicaux, chacun avec le résumé de l'auteur correspondant, 124.929 mots ;
6. Résumé générique mono-document en français, revue *Perspectives interdisciplinaires sur le travail et la santé* (PISTES)[6] ; 50 articles et leurs résumés des auteurs, 381.039 mots ;
7. Résumé générique multi-document en français, corpus RPM2[7] (journalistique) ; 20 thématiques différentes composées de 10 articles et 4 résumés de référence par thématique, 185.223 mots.

Pour les expériences avec les corpora TAC et DUC nous avons utilisé directement les résumés candidats produits par les systèmes participant aux évaluations (données officielles). Pour les expériences en espagnol et en français (résumé mono et multi-document), nous avons créé des résumés à un taux de compression similaire à ceux de référence en utilisant les systèmes suivants :

– *ENERTEX* (Fernandez *et al.*, 2007), un système de résumé automatique fondée sur l'énergie textuelle ;
– *CORTEX* (Torres-Moreno *et al.*, 2002), un système d'extraction de phrases mono-document multilangue qui combine différentes mesures statistiques de pertinence (angle entre les phrases et la thématique, le poids de Hamming des phrases, etc.) et applique un algorithme de décision optimale pour la sélection des phrases pertinentes ;

[2] Récupérable à l'adresse : `http://lia.univ-avignon.fr/fileadmin/axes/TALNE/Ressources.html`
[3] `http://www-nlpir.nist.gov/projects/duc/guidelines/2004.html`
[4] `http://www.nist.gov/tac/data/index.html`
[5] `http://www.elsevier.es/revistas/ctl_servlet?_f=7032&revistaid=2`
[6] `http://www.pistes.uqam.ca/`
[7] `http://labs.sinequa.com/rpm2/`

– *SUMMTERM* (Vivaldi *et al.*, 2010), système de résumé des articles médicaux basé sur la terminologie spécialisée afin de donner un score et un classement aux phrases ;
– *REG* (Torres-Moreno & Ramirez, 2010), système de résumé basé sur un algorithme glouton ;
– Résumeur $\mathcal{JS}$, système qui donne un score et un classement aux phrases en considérant leur divergence Jensen-Shannon par rapport au document source ;
– *Baseline-premières phrases*, qui sélectionne les premières phrases du document pour construire les résumés ;
– *Baseline-aléatoire*, sélectionne les phrases au hasard pour construire les résumés ;
– *Open Text Summarizer* (Yatsko & Vishnyakov, 2007), résumeur multi-langue basé sur la fréquence, et
– les systèmes de résumé commerciaux multi-langues *Word*, *SSSummarizer*[8], *Pertinence*[9] et *Copernic*[10].

3.3 Mesures d'évaluation

Les mesures suivantes, qui proviennent de l'évaluation humaine du contenu des résumés, ont été utilisées dans nos expériences :
– Couverture : la quantité d'information partagée entre un résumé candidat et un résumé de référence (Over *et al.*, 2007). Elle a été utilisée dans les campagnes d'évaluation DUC.
– *Responsiveness* : elle classe les résumés sur une échelle de 5, en indiquant dans quelle mesure le résumé répond à un besoin d'information précis (Over *et al.*, 2007). Elle est utilisée dans les tâches de résumé guidé, comme cela a été le cas de certaines tâches des campagnes DUC et TAC.
– PYRAMIDS : elle vérifie que les unités d'information essentielles (telles qu'on les trouve dans des résumés de référence générés par les humains) soient présents dans les résumés candidats. PYRAMIDS est la mesure d'évaluation basée sur le contenu des campagnes TAC.

Pour les corpus DUC et TAC, les valeurs de ces mesures sont disponibles et nous les avons utilisées directement. Dans nos expériences nous avons utilisé les mesures d'évaluation automatiques suivantes :
– ROUGE : métrique de rappel qui emploie des n-grammes comme unités de contenu pour comparer les résumés candidats vs. ceux de référence. L'équation ROUGE spécifiée dans (Lin, 2004) est la suivante :

$$\text{ROUGE-}n(\text{R}, M) = \frac{\sum_m \in M \sum_{n-\text{gramme} \in P} \text{count}_{\text{match}}(n-\text{gramme})}{\sum_m \in M \sum \text{count(n-gramme)}} \tag{4}$$

où R est le résumé à évaluer, M est l'ensemble des résumés modèles (humains), $\text{count}_{\text{match}}$ le nombre de n-grammes communs en m et P, et count est le nombre de n-grammes dans les résumés modèles. Pour nos expériences, nous avons utilisé uni-grammes, 2-grammes et *skip* 2-grammes avec une distance maximale de 4 (ROUGE-1, ROUGE-2 et ROUGE-SU4). ROUGE est utilisée pour comparer un résumé candidat à l'ensemble des résumés de référence disponibles.
– L'équation 1 de la divergence de $\mathcal{JS}$ a été implémentée dans notre système FRESA avec la spécification suivante pour la distribution de probabilités de mots w :

$$P_w = \frac{C_w^T}{N}; \qquad Q_w = \begin{cases} \frac{C_w^S}{N_S} & \text{si } w \in S \\ \frac{C_w^T + \delta}{N + \delta * B} & \text{autrement} \end{cases} \tag{5}$$

où P est la distribution de probabilités des mots w dans le texte T et Q la distribution de probabilités des mots w dans le résumé S ; N est le nombre de mots dans le texte et le résumé $N = N_T + N_S$,

[8] http://www.kryltech.com/summarizer.htm
[9] http://www.pertinence.net
[10] http://www.copernic.com/en/products/summarizer

$B = 1.5|V|$, C_w^T est le nombre de mots dans le texte et C_w^S est le nombre de mots dans le résumé. Pour le lissage des probabilités du résumé, nous avons utilisé $\delta = 0.005$. Nous avons également implémenté d'autres méthodes de lissage (par exemple, Good-Turing Manning & Schütze (1999) qui utilise le package statistique de Perl *Statistics-Smoothing-SGT-2.1.2*[11]) dans FRESA, mais nous ne les utilisons pas dans les expériences rapportées ici. À l'instar de l'approche ROUGE, en plus des uni-grammes de mots nous avons utilisé des 2-grammes et *skip* 2-grammes pour le calcul des divergences telles que $\mathcal{JS}$ (à l'aide des uni-grammes), $\mathcal{JS}_2$ (à l'aide des 2-grammes), $\mathcal{JS}_4$ (à l'aide des *skip* 2-grammes de ROUGE-SU4) et $\mathcal{JS}_M$ qui est la moyenne des $\mathcal{JS}_i$. Les mesures $\mathcal{JS}$ sont utilisées pour comparer les résumés candidats à son document(s) source, dans notre cadre.

4 Expériences et résultats

Dan premier temps, nous avons reproduit les expériences présentées dans Louis & Nenkova (2009) pour vérifier que notre implémentation $\mathcal{JS}$ obtient des résultats de corrélation cohérents avec ce travail. Nous avons utilisé les corpus de résumés mis à jour de TAC'08 pour calculer les mesures $\mathcal{JS}$ et ROUGE pour chaque résumé candidat. Nous avons réalisé deux classements des systèmes (un pour chaque mesure), qui ont été comparés aux classements produits selon les scores de PYRAMIDS et de *Responsiveness*. Les corrélations de Spearman ont été calculées entre les différents classements. Les résultats sont présentés au tableau 1 avec leur p-value[12]. Ils confirment une forte corrélation entre PYRAMIDS, *Responsiveness* et $\mathcal{JS}$. Nous avons également vérifié la corrélation élevée entre $\mathcal{JS}$ et ROUGE-2 ($0,83$ Spearman, non affichée dans le tableau) pour cette tâche et ce corpus. Puis, nous avons mené des expériences sur les corpus

Mesure	PYRAMIDS	p-value	*Responsiveness*	p-value
ROUGE-2	0,96	$p < 0,005$	0,92	$p < 0,005$
$\mathcal{JS}$	0,85	$p < 0,005$	0,74	$p < 0,005$

TAB. 1 – Corrélation de Spearman, mesures d'informativité, TAC'08 *Update Summarization*

DUC'04 et TAC'08 pour la tâche pilote de résumés d'opinion. Nous avons aussi mené des expériences de résumé mono et multi-document en français et espagnol. En dépit du fait que les expériences pour les corpus français et espagnol utilisent moins de systèmes ou de documents (par exemple, un nombre inférieur de résumés par tâche) que pour l'anglais, les résultats restent significatifs. Pour DUC'04, nous avons calculé la mesure $\mathcal{JS}$ pour chaque résumé candidat des tâches 2 et 5 et nous avons calculer les différents classements des systèmes induits par $\mathcal{JS}$, ROUGE, les score de Couverture et *Responsiveness*. Les différentes valeurs de la corrélation de classement Spearman pour DUC'04 sont présentées aux tableaux 2 (pour la tâche 2) et 3 (pour la tâche 5). Pour la tâche 2, nous avons vérifié une forte corrélation entre $\mathcal{JS}$ et la Couverture. Pour la tâche 5, la corrélation entre $\mathcal{JS}$ et la Couverture est faible, et la corrélation entre $\mathcal{JS}$ et *Responsiveness* est faible voire négative. Bien que la tâche de résumé d'opinion soit récente et son évaluation un problème compliqué, nous avons décidé de comparer les classements $\mathcal{JS}$ avec ceux de PYRAMIDS et *Responsiveness* sur le corpus de TAC'08. Les valeurs de la corrélation de Spearman sont affichées au tableau 4. Comme on peut le constater, il existe une corrélation faible voire négative entre $\mathcal{JS}$ et PYRAMIDS ou *Responsiveness*. La corrélation entre les classement PYRAMIDS et *Responsiveness*

[11] CPAN, `http://search.cpan.org/~bjoernw/Statistics-Smoothing-SGT-2.1.2/`

[12] En statistique, la p-value est le plus petit niveau auquel on rejette l'hypothèse nulle.

Mesure	Couverture	p-value
ROUGE-2	0,79	$p < 0,0050$
$\mathcal{JS}$	0,68	$p < 0,0025$

TAB. 2 – Corrélation de Spearman, mesures $\mathcal{JS}$ et ROUGE vs. Couverture, DUC'04 tâche 2

Mesure	Couverture	p-value	*Responsiveness*	p-value
ROUGE-2	0,78	$p < 0,001$	0,44	$p < 0,05$
$\mathcal{JS}$	0,40	$p < 0,050$	-0,18	$p < 0,25$

TAB. 3 – Corrélation de Spearman, $\mathcal{JS}$ et ROUGE vs. *Responsiveness* et Couverture, DUC'04 tâche 5

est élevée pour cette tâche (0,71 valeur de corrélation de Spearman). Pour les expériences en espagnol et

Mesure	PYRAMIDS	p-value	*Responsiveness*	p-value
$\mathcal{JS}$	-0,13	$p < 0,25$	-0,14	$p < 0,25$

TAB. 4 – Corrélation de Spearman, mesure $\mathcal{JS}$ vs. *Responsiveness* et PYRAMIDS, TAC'08 OS

en français mono-document, nous avons lancé 11 systèmes de résumé multi-langue sur les corpus. Pour l'expérience en français multi-document nous avons utilisé 12 systèmes. Dans tous les cas, nous avons produit des résumés aux taux de compression proches de ceux des résumés des auteurs (*abstracts*). Puis nous avons calculé les mesures $\mathcal{JS}$ et ROUGE pour chaque résumé et nous avons obtenu la moyenne des valeurs pour chaque système. Ces moyennes ont été utilisées pour produire les classements pour chaque mesure. Nous avons calculé les corrélations de Spearman pour toutes les paires de classement. Les résultats sont présentés dans les tablaux 5, 6 et 7. Ils montrent une corrélation moyenne à forte entre les mesures $\mathcal{JS}$ et celles ROUGE. Toutefois, la mesure $\mathcal{JS}$ basée sur les uni-grammes obtient une corrélation inférieure à celle qui utilise des n-grammes d'ordre supérieur.

5 Discussion

Nos recherches s'inspirent des récents travaux sur l'utilisation des métriques d'évaluation basées sur le contenu qui ne reposent pas sur des références humaines, mais qui comparent le contenu des résumés directement aux entrées (Louis & Nenkova, 2009). Nous avons obtenu des résultats positifs et négatifs en ce qui concerne l'utilisation directe des documents à résumer pour établir des mesures d'évaluation par contenu. Nous avons vérifié que dans les deux variétés de résumé multi-document en anglais, générique et basé sur une thématique, la corrélation entre les mesures qui utilisent de références humaines (PYRAMIDS, *Responsiveness* et ROUGE) et une mesure qui n'utilise pas de référence (la divergence de $\mathcal{JS}$) est forte. Nous avons trouvé que la corrélation entre les mêmes mesures est faible pour le résumé d'informations biographiques et le résumé d'opinions des blogs. Nous pensons que dans ces cas, les mesures fondées sur le contenu devraient prendre en compte en plus du document d'entrée, la tâche du résumé (par exemple la représentation texte de la tâche-question, description, etc.) pour mieux évaluer le contenu des candidats (Spärck Jones, 2007), puisque la tâche est un facteur déterminant dans la sélection du contenu pour

Mesure	ROUGE-1	p-value	ROUGE-2	p-value	ROUGE-SU4	p-value
$\mathcal{JS}$	0,56	$p < 0,100$	0,46	$p < 0,100$	0,45	$p < 0,200$
$\mathcal{JS}_2$	0,88	$p < 0,001$	0,80	$p < 0,002$	0,81	$p < 0,005$
$\mathcal{JS}_4$	0,88	$p < 0,001$	0,80	$p < 0,002$	0,81	$p < 0,005$
$\mathcal{JS}_M$	0,82	$p < 0,005$	0,71	$p < 0,020$	0,71	$p < 0,010$

TAB. 5 – Corrélation de Spearman, $\mathcal{JS}$ vs. ROUGE, corpus mono-document *Medicina Clínica* (espagnol)

Mesure	ROUGE-1	p-value	ROUGE-2	p-value	ROUGE-SU4	p-value
$\mathcal{JS}$	0,70	$p < 0,050$	0,73	$p < 0,05$	0,73	$p < 0,500$
$\mathcal{JS}_2$	0,93	$p < 0,002$	0,86	$p < 0,01$	0,86	$p < 0,005$
$\mathcal{JS}_4$	0,83	$p < 0,020$	0,76	$p < 0,05$	0,76	$p < 0,050$
$\mathcal{JS}_M$	0,88	$p < 0,010$	0,83	$p < 0,02$	0,83	$p < 0,010$

TAB. 6 – Corrélation de Spearman, mesures $\mathcal{JS}$ vs. ROUGE, corpus mono-document PISTES (français)

le résumé. Les expériences multi-langue réalisées en résumé générique mono-document confirment une

Mesure	ROUGE-1	p-value	ROUGE-2	p-value	ROUGE-SU4	p-value
$\mathcal{JS}$	0,830	$p < 0,002$	0,660	$p < 0,05$	0,741	$p < 0,01$
$\mathcal{JS}_2$	0,800	$p < 0,005$	0,590	$p < 0,05$	0,680	$p < 0,02$
$\mathcal{JS}_4$	0,750	$p < 0,010$	0,520	$p < 0,10$	0,620	$p < 0,05$
$\mathcal{JS}_M$	0,850	$p < 0,002$	0,640	$p < 0,05$	0,740	$p < 0,01$

TAB. 7 – Corrélation de Spearman, mesures $\mathcal{JS}$ vs. ROUGE, corpus multi-document RPM2 (français)

forte corrélation entre les mesures de divergence $\mathcal{JS}$ et ROUGE. Il faut noter que ROUGE est en général le logiciel choisi pour présenter les résultats des évaluations basées sur le contenu des résumés en d'autres langues que l'anglais. Pour les expériences en espagnol, nous sommes conscients que nous n'avons qu'un seul résumé de référence à comparer avec les résumés candidats. Néanmoins, ces références sont les résumés d'auteurs. Comme le montrent les expériences conduites par da Cunha *et al.* (2007), les professionnels d'un domaine spécialisé (par exemple le domaine médical) adoptent des stratégies similaires pour résumer leurs textes : ils ont tendance à choisir des passages à peu près du même contenu pour leurs résumés. Des études antérieures ont montré qu'à partir des résumés d'auteur, il est possible de reformuler fidèlement le contenu du texte (Chuah, 2001). De ce fait, le résumé de l'auteur d'un article médical peut être pris comme référence pour l'évaluation des résumés. Dans le corpus en français PISTES, on suppose une situation semblable au cas en espagnol.

6 Conclusions et travail futur

Dans cet article, nous avons étudié la validité des mesures d'évaluation du contenu des résumés sans utilisation de résumés de référence. Il est à noter qu'il y a un débat sur le nombre de références à utiliser pour évaluer les résumés (Owkzarzak & Dang, 2009). Nous avons mené de multiples expérimentations sur

un large spectre de tâches du résumé mono-document générique au résumé des opinions. Les principales contributions de cet article sont les suivantes :

– Nous avons montré que, si l'on s'intéresse uniquement au classement des résumeurs par informativité, il y a des tâches où les références pourraient être substituées par le document complet, tout en obtenant des classements fiables. Cependant, nous avons aussi constaté que la substitution des références par le document complet n'est pas toujours souhaitable. Nous n'avons ainsi trouvé qu'une faible corrélation entre les différents classements dans des tâches de résumé complexes telles que le résumé biographique et le résumé d'opinion.
– Nous avons également effectué des expériences à grande échelle en espagnol et en français qui montrent une corrélation positive de moyenne à forte entre les classements des systèmes obtenus par ROUGE et les mesures de divergence qui n'utilisent pas des résumés de référence.
– Nous avons présenté les résultats du système FRESA, pour le calcul des mesures basées sur la divergence $\mathcal{JS}$. De même que dans ROUGE, FRESA utilise des uni-grammes, 2-grammes et des *skip* 2-grammes de mots pour le calcul des divergences.

Bien d'autres développements et expérimentations sont envisageables. Ainsi, afin de vérifier la corrélation entre ROUGE et $\mathcal{JS}$, à court terme, nous avons l'intention d'étendre nos recherches à d'autres langues et corpora telles que le portugais et le chinois pour lesquels nous avons accès aux données et à la technologie pour produire des résumés dans plusieurs langues. Nous envisageons également d'appliquer FRESA aux autres tâches de résumé de DUC et TAC. À plus long terme, nous prévoyons d'intégrer une représentation de la tâche/thématique dans le calcul des mesures. Pour mener à bien toutes ces comparaisons nous sommes cependant dépendants de l'existence de références, sans lesquelles on ne peut mesurer la proximité entre résumés automatiques et performances humaines. Enfin, FRESA sera utilisé dans la nouvelle tâche de question-réponse (QA) de la campagne INEX (`http://www.inex.otago.ac.nz/tracks/qa/qa.asp`) pour l'évaluation des réponses longues. Cette tâche consiste à répondre à une question de type encyclopédique par extraction et agglomération de phrases de Wikipédia. Ce type de texte et de question correspond bien à ceux des tâches pour lesquelles nous avons constaté des taux de corrélation élevés entre les mesures $\mathcal{JS}$ et les méthodes d'évaluation avec intervention humaine. Par ailleurs, le calcul de la divergence se fera entre les résumés produits et un ensemble représentatif des passages pertinents du Wikipédia. FRESA sera ainsi utilisé pour comparer trois types de systèmes appliqués à une même tâche : les résumeurs multi-documents guidés par une requête, les systèmes de recherche d'information ciblées (focused IR) et les systèmes de QA.

Remerciements. Ce travail a été financé partiellement par la bourse post-doctorale d'Iria da Cunha (*Ministerio Español de Ciencia e Innovación*, MICINN). H. Saggion remercie les supports des programmes Ramon y Cajal (MICINN) et Começa 2010 de l'Universitat Pompeu Fabra, Barcelone.

Références

CHUAH C.-K. (2001). Types of lexical substitution in abstracting. In *ACL Student Research Workshop*, p. 49–54, Toulouse, France : Association for Computational Linguistics.

DA CUNHA I., WANNER L. & CABRÉ M. T. (2007). Summarization of specialized discourse : The case of medical articles in spanish. *Terminology*, **13**(2), 249–286.

DONAWAY R. L., DRUMMEY K. W. & MATHER L. A. (2000). A comparison of rankings produced by summarization evaluation measures. In *NAACL Workshop on Automatic Summarization*, p. 69–78.

FERNANDEZ S., SANJUAN E. & TORRES-MORENO J.-M. (2007). Textual Energy of Associative Memories : performants applications of Enertex algorithm in text summarization and topic segmentation. In *MICAI'07*, p. 861–871.

KULLBACK S. & LEIBLER R. (1951). On information and sufficiency. *Ann. of Math. Stat.*, **22**(1), 79–86.

LIN C.-Y. (2004). ROUGE : A Package for Automatic Evaluation of Summaries. In M.-F. MOENS & S. SZPAKOWICZ, Eds., *Text Summarization Branches Out : ACL-04 Workshop*, p. 74–81, Barcelona.

LIN C.-Y., CAO G., GAO J. & NIE J.-Y. (2006). An information-theoretic approach to automatic evaluation of summaries. In *HLT-NAACL*, p. 463–470, Morristown, USA.

LIN J. (1991). Divergence Measures based on the Shannon Entropy. *IEEE Tr. on Inf. Th.*, **37**(145-151).

LOUIS A. & NENKOVA A. (2009). Automatically Evaluating Content Selection in Summarization without Human Models. In *Empirical Methods in Natural Language Processing*, p. 306–314, Singapore.

MANI I., KLEIN G., HOUSE D., HIRSCHMAN L., FIRMIN T. & SUNDHEIM B. (2002). Summac : a text summarization evaluation. *Natural Language Engineering*, **8**(1), 43–68.

MANNING C. D. & SCHÜTZE H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts : The MIT Press.

NENKOVA A. & PASSONNEAU R. J. (2004). Evaluating Content Selection in Summarization : The Pyramid Method. In *HLT-NAACL*, p. 145–152.

OVER P., DANG H. & HARMAN D. (2007). DUC in context. *IPM*, **43**(6), 1506–1520.

OWKZARZAK K. & DANG H. T. (2009). Evaluation of automatic summaries : Metrics under varying data conditions. In *UCNLG+Sum'09*, p. 23–30, Suntec, Singapore.

PAPINENI K., ROUKOS S., WARD T., & ZHU W. J. (2002). BLEU : a method for automatic evaluation of machine translation. In *ACL'02*, p. 311–318.

PASTRA K. & SAGGION H. (2003). Colouring summaries BLEU. In *Evaluation Initiatives in Natural Language Processing*, Budapest, Hungary : EACL.

RADEV D. R., TEUFEL S., SAGGION H., LAM W., BLITZER J., QI H., ÇELEBI A., LIU D. & DRÁBEK E. (2003). Evaluation challenges in large-scale document summarization. In *ACL'03*, p. 375–382.

SAGGION H., RADEV D., TEUFEL S. & LAM W. (2002). Meta-evaluation of Summaries in a Cross-lingual Environment using Content-based Metrics. In *COLING 2002*, p. 849–855, Taipei, Taiwan.

SIEGEL S. & CASTELLAN N. (1998). *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill.

SPÄRCK JONES K. (2007). Automatic summarising : The state of the art. *IPM*, **43**(6), 1449–1481.

SPÄRCK JONES K. & GALLIERS J. (1996). *Evaluating Natural Language Processing Systems, An Analysis and Review*, volume 1083 of *Lecture Notes in Computer Science*. Springer.

TAC (2008). *Proceedings of the Text Analysis Conference*, Gaithesburg, Maryland, USA. NIST.

TORRES-MORENO J.-M. & RAMIREZ J. (2010). REG : un algorithme glouton appliqué au résumé automatique de texte. In *JADT'10* : Rome.

TORRES-MORENO J.-M., VELÁZQUEZ-MORALES P. & MEUNIER J.-G. (2002). Condensés de textes par des méthodes numériques. In *JADT'02*, volume 2, p. 723–734, St Malo, France.

VIVALDI J., DA CUNHA I., TORRES-MORENO J.-M. & VELÁZQUEZ-MORALES P. (2010). Automatic summarization using terminological and semantic resources. In *LREC'10*, Malta.

YATSKO V. & VISHNYAKOV T. (2007). A method for evaluating modern systems of automatic text summarization. *Automatic Documentation and Mathematical Linguistics*, **41**(3), 93–103.